CLAIMS

What is claimed is:

1. A method for identifying a small subgroup of compounds representative of a larger set of compounds, said method comprising:

providing a set of compounds;

obtaining one or more descriptor values for each compound in the set of compounds;

determining a median value for each of the descriptor values for the set of compounds;

partitioning the set of compounds into a plurality of partitions using each median value for the set of compounds; and

selecting compounds from each of the plurality of partitions to form a subgroup of compounds representative of the set of compounds.

2. The method as set forth in claim 1 further comprising:

repeating said obtaining, determining, and partitioning one or more times with different descriptor values than used previously.

3. The method as set forth in claim 1, wherein said partitioning the compounds into partitions comprises:

dividing the compounds into a first partition of compounds which have the descriptor value greater than the median value and a second partition which have the descriptor value less than the median value.

4. The method as set forth in claim 1, wherein said selecting comprises:

determining a partition median value for each of the descriptor values for the compounds within a partition; and

selecting from the partition one or more compounds that have each descriptor value being within a predetermined range of values away from a corresponding partition median value to represent the compounds within the partition.

5. The method as set forth in claim 1, wherein the descriptor values are descriptor types independently selected from the group consisting of chemical properties, structural properties, surface area properties, and electrochemical properties.

6. The method as set forth in claim 1, wherein the descriptor values are descriptor types independently selected from the group consisting of a sum of atomic polarizabilities of all atoms, a number of aromatic atoms, a number of H-bond donors, a number of heavy atoms, a number of hydrophobic atoms, a number of nitrogen atoms, a number of fluorine atoms, a number of sulfur atoms, a number of iodine atoms, a number of bonds between heavy atoms, a number of aromatic bonds, a number of double nonaromatic bonds, an atomic connectivity index (order 0), a carbon valence connectivity index (order 1), a carbon connectivity index (order 1), a greatest value in a distance matrix, a third kappa shape index, a relative negative partial charge, a total positive van der Waals surface area, a fractional negative polar van der Waals surface area, a fractional hydrophobic van der Waals surface area, a vertex adjacency information (magnitude), a vertex distance equality index, a vertex distance magnitude index, a sum of a van der Waals surface area of each of one or more atoms in each compound in the set of compounds, a van der Waals surface area calculated for a property of each compound selected from the group consisting of hydrogen-bond acceptor atoms, hydrogen-bond donor atoms, nondonor-acceptor atoms, and polar atoms, a van der Waals volume calculated using a connection table, and a Zagreb index.

7. The method as set forth in claim 1, wherein the descriptor values are different descriptor types that do not substantially correlate with each other.

8. The method as set forth in claim 1, further comprising:

choosing different types of descriptors to base the descriptor values on using a genetic algorithm.

9. The method as set forth in claim 8, wherein the different types of descriptors for the set of compounds each have value distributions from which the median values are calculated.

10. The method as set forth in claim 8 further comprising:

establishing an optimal combination of the different types of descriptors to base the descriptor values on using the genetic algorithm.

11. The method as set forth in claim 10, wherein a scoring function is used by the genetic algorithm during said establishing of the optimal combination of the different types of descriptors, the scoring function comprising:

$$S = \frac{100}{N_{total}} \times \frac{1}{(N_{total} - N_P) + C/C_{act}},$$

wherein N_{total} is a first total number of active compounds in the set of compounds, N_P is a second total number of compounds in partitions which have one type of compound, C is a third total number of partitions which have one or more types of compounds, and C_{act} is a fourth total number of one or more activity classes present in the set of compounds.

12. The method as set forth in claim 1, wherein said obtaining one or more descriptor values comprises:

calculating the descriptor values using a molecular modeling program.

13. A computer-readable medium having stored thereon instructions for identifying a small subgroup of compounds representative of a larger set of compounds, which when executed by at least one processor, causes the processor to perform:

providing information representing a set of compounds;

obtaining one or more descriptor values for each compound in the set of compounds;

determining a median value for each of the descriptor values for the set of compounds;

partitioning the set of compounds into a plurality of partitions using each median value for the set of compounds; and

selecting compounds from each of the plurality of partitions to form a subgroup of compounds representative of the set of compounds.

14. The medium as set forth in claim 13 further comprising:

repeating said obtaining, determining, and partitioning one or more times with different descriptor values than used previously.

15. The medium as set forth in claim 13, wherein said partitioning the compounds into partitions comprises:

dividing the compounds into a first partition of compounds which have the descriptor value greater than the median value and a second partition which have the descriptor value less than the median value.

16. The medium as set forth in claim 13, wherein said selecting comprises:

determining a partition median value for each of the descriptor values for the compounds within a partition; and

selecting from the partition one or more compounds that have each descriptor value being within a predetermined range of values away from a corresponding partition median value to represent the compounds within the partition.

17. The medium as set forth in claim 13, wherein the descriptor values are descriptor types independently selected from the group consisting of chemical properties, structural properties, surface area properties, and electrochemical properties.

18. The medium as set forth in claim 13, wherein the descriptor values are descriptor types independently selected from the group consisting of a sum of atomic polarizabilities of all atoms, a number of aromatic atoms, a number of H-bond donors, a number of heavy atoms, a number of hydrophobic atoms, a number of nitrogen atoms, a number of fluorine atoms, a number of sulfur atoms, a number of iodine atoms, a number of bonds between heavy atoms, a number of aromatic bonds, a number of double nonaromatic bonds, an atomic connectivity index (order 0), a carbon valence connectivity index (order 1), a carbon connectivity index (order 1), a greatest value in a distance matrix, a third kappa shape index, a relative negative partial charge, a total positive van der Waals surface area, a fractional negative polar van der Waals surface area, a fractional hydrophobic van der Waals surface area, a vertex adjacency information (magnitude), a vertex distance equality index, a sum of a van der Waals surface area of each of one or more atoms in each compound in the set of compounds, a van der Waals surface area calculated for a property of each compound selected from the group consisting of hydrogen-bond acceptor atoms, hydrogen-bond donor atoms, nondonor-acceptor atoms, and polar atoms, a vertex distance magnitude index, a van der Waals volume calculated using a connection table, and a Zagreb index.

19. The medium as set forth in claim 13, wherein the descriptor values are different descriptor types that do not substantially correlate with each other.

20. The medium as set forth in claim 13 further comprising:

choosing different types of descriptors to base the descriptor values on using a genetic algorithm.

21. The medium as set forth in claim 20 wherein the different types of descriptors for the set of compounds each have value distributions from which the median values are calculated.

22. The medium as set forth in claim 20, further comprising:

establishing an optimal combination of the different types of descriptors to base the descriptor values on using the genetic algorithm.

23. The medium as set forth in claim 22, wherein a scoring function is used by the genetic algorithm during said establishing of the optimal combination of the different types of descriptors, the scoring function comprising:

$$S = \frac{100}{N_{total}} \times \frac{1}{(N_{total} - N_P) + C/C_{act}},$$

wherein N_{total} is a first total number of active compounds in the set of compounds, N_P is a second total number of compounds in partitions which have one type of compound, C is a third total number of partitions which have one or more types of compounds, and C_{act} is a fourth total number of one or more activity classes present in the set of compounds.

24. The medium as set forth in claim 13, wherein said obtaining one or more descriptor values comprises:

calculating the descriptor values using a molecular modeling program.

25. A system for identifying a small group of compounds representative of a larger set of compounds, said system comprising:

a descriptor system that obtains one or more descriptor values for information representing each compound in the set of compounds;

a median determination system that determines a median value for each of the descriptor values for the set of compounds;

a partitioning system that partitions the set of compounds into a plurality of partitions using each median value for the set of compounds; and

a partition selection system that selects compounds from each of the plurality of partitions to form a subgroup representative of the set of compounds.

26. The system as set forth in claim 25, wherein the partition selection system causes operation of the descriptor system, the median determination system, and the partitioning system one or more times, the descriptor values each being a different type of descriptor than the descriptor values used previously.

27. The system as set forth in claim 25, wherein the partitioning system divides the compounds into a first partition of compounds which have the descriptor value greater than the median value and a second partition which have the descriptor value less than the median value.

28. The system as set forth in claim 25, wherein the partition selection system determines a partition median value for each of the descriptor values for the compounds within a partition and selects from the partition one or more compounds that have each descriptor value being within a predetermined range of values away from a corresponding partition median value to represent the compounds within the partition.

29. The system as set forth in claim 25, wherein the descriptor values are different descriptor types that do not substantially correlate with each other.

30. The system as set forth in claim 25, wherein the descriptor system chooses different types of descriptors to base the descriptor values on using a genetic algorithm.

31. The system as set forth in claim 30, wherein the different types of descriptors for the set of compounds each have value distributions from which the median values are calculated.

32. The system as set forth in claim 30, wherein the descriptor system establishes an optimal combination of the different types of descriptors to base the descriptor values on using the genetic algorithm.

33. The system as set forth in claim 32, wherein a scoring function is used by the genetic algorithm during establishment of the optimal combination of the different types of descriptors, the scoring function comprising:

$$S = \frac{100}{N_{total}} \times \frac{1}{(N_{total} - N_P) + C/C_{act}},$$

wherein N_{total} is a first total number of active compounds in the set of compounds, N_P is a second total number of compounds in partitions which have one type of compound, C is a third total number of partitions which have one or more types of compounds, and C_{act} is a fourth total number of one or more activity classes present in the set of compounds.

34. The system as set forth in claim 25, wherein the descriptor system calculates the descriptor values using a molecular modeling program.

35. A method for virtual compound screening comprising:

combining a plurality of unidentified compounds with a plurality of bait compounds with known biological activities to create a set of compounds;

obtaining one or more descriptor values for each of the unidentified compounds and for each of the bait compounds in the set of compounds;

determining a median value for each of the descriptor values for the set of compounds;

partitioning the set of compounds into a plurality of partitions based on each median value;

recombining partitions which have at least two bait compounds to form a recombined set of compounds; and

selecting the recombined set of compounds for analysis of biological activity if an approximate target number of unidentified components remain in the recombined set of compounds.

36. The method as set forth in claim 35 further comprising:

repeating said obtaining, determining, partitioning and recombining with different descriptor values than used previously until the approximate target number of unidentified compounds remain in the recombined set of compounds.

37. The method as set forth in claim 36 further comprising:

reintroducing another set of bait compounds into the recombined set of compounds substantially prior to repeating said obtaining, the other set of bait compounds are identical to the bait compounds used during said combining.

38. The method as set forth in claim 35, wherein the target number of compounds is less than about 100 compounds.

39. The method as set forth in claim 35, wherein each bait compound comprises an active compound selected from the group consisting of benzodiazepine receptor ligands, serotonin receptor ligands, tyrosine kinase inhibitors, histamine H3 antagonists, cyclooxygenase-2 inhibitors, HIV protease inhibitors, carbonic anhydrase II inhibitors, β-lactamase inhibitors, protein kinase C inhibitors, estrogen antagonists, antihypertensive (ACE inhibitor), antiadrenergic (β-receptor), glucocorticoid analogues, angiotensin AT1 antagonists, aromatase inhibitors, DNA topoisomerase I inhibitors, dihydrofolate reductase inhibitors, factor Xa inhibitors, farnesyl transferase inhibitors, matrix metalloproteinase inhibitors, and vitamin D analogues.

40. The method as set forth in claim 35, wherein each bait compound has a particular biological activity.

41. The method as set forth in claim 35, wherein said partitioning the compounds into partitions comprises:

dividing the compounds into a first partition of compounds which have the descriptor value greater than the median value and a second partition which have the descriptor value less than the median value.

42. The method as set forth in claim 35, wherein the descriptor values are different descriptor types independently selected from the group consisting of chemical properties, structural properties, surface area properties, and electrochemical properties.

43. The method as set forth in claim 35, wherein the descriptor values are descriptor types independently selected from the group consisting of a sum of atomic polarizabilities of all atoms, a number of aromatic atoms, a number of H-bond donors, a number of heavy atoms, a number of hydrophobic atoms, a number of nitrogen atoms, a number of fluorine atoms, a number of sulfur atoms, a number of iodine atoms, a number of bonds between heavy atoms, a number of aromatic bonds, a number of double nonaromatic bonds, an atomic connectivity index (order 0), a carbon valence connectivity index (order 1), a carbon connectivity index (order 1), a greatest value in a distance matrix, a third kappa shape index, a relative negative partial charge, a total positive van der Waals surface area, a fractional negative polar van der Waals surface area, a fractional hydrophobic van der Waals surface area, a vertex adjacency information (magnitude), a vertex distance equality index, a vertex distance magnitude index, a sum of a van der Waals surface area of each of one or more atoms in each compound in the set of compounds, a van der Waals surface area calculated for a property of each compound selected from the group consisting of hydrogen-bond acceptor atoms, hydrogen-bond donor atoms, nondonor-acceptor atoms, and polar atoms, a van der Waals volume calculated using a connection table, and a Zagreb index.

44. The method as set forth in claim 35, wherein the descriptor values are different descriptor types that do not substantially correlate with each other.

45. The method as set forth in claim 35 further comprising:

choosing different types of descriptors to base the descriptor values on using a genetic algorithm.

46. The method as set forth in claim 45, wherein the different types of descriptors for the set of compounds each have value distributions from which the median values are calculated.

47. The method as set forth in claim 45 further comprising:

establishing an optimal combination of the different types of descriptors to base the descriptor values on using the genetic algorithm.

48. The method as set forth in claim 45, wherein a scoring function is used by the genetic algorithm during said establishing of the optimal combination of the different types of descriptors, the scoring function comprising:

$$S = Act(cp) \times Pa(pop),$$

wherein *Act(cp)* is a first total number of co-partitioned known active compounds in the set of compounds and *Pa(pop)* is a second total number of populated partitions.

49. The method as set forth in claim 35, wherein said obtaining one or more descriptor values comprises:

calculating the descriptor values using a molecular modeling program.

50. A computer-readable medium having stored thereon instructions for virtual compound screening, which when executed by at least one processor, causes the processor to perform:

combining information representing a plurality of unidentified compounds with information representing a plurality of bait compounds with known biological activities to create a set of compounds;

obtaining one or more descriptor values for each of the unidentified compounds and for each of the bait compounds in the set of compounds;

determining a median value for each of the descriptor values for the set of compounds;

partitioning the set of compounds into a plurality of partitions based on each median value;

recombining partitions which have at least two bait compounds to form a recombined set of compounds; and

selecting the recombined set of compounds for analysis of biological activity if an approximate target number of unidentified compounds remain in the recombined set of compounds.

51. The medium as set forth in claim 50 comprising:

repeating said obtaining, determining, partitioning and recombining with different descriptor values than used previously until the approximate target number of unidentified compounds remain in the recombined set of compounds.

52. The medium as set forth in claim 51 further comprising:

reintroducing another set of bait compounds into the recombined set of compounds substantially prior to repeating said obtaining, the other set of bait compounds are identical to the bait compounds used during said combining.

53. The medium as set forth in claim 50, wherein the target number of compounds is less than about 100 compounds.

54. The medium as set forth in claim 50, wherein each bait compound comprises an active compound selected from the group consisting of benzodiazepine receptor ligands, serotonin receptor ligands, tyrosine kinase inhibitors, histamine H3 antagonists, cyclooxygenase-2 inhibitors, HIV protease inhibitors, carbonic anhydrase II inhibitors, β-lactamase inhibitors, protein kinase C inhibitors, estrogen antagonists, antihypertensive (ACE inhibitor), antiadrenergic (β-receptor), glucocorticoid analogues, angiotensin AT1 antagonists, aromatase inhibitors, DNA topoisomerase I inhibitors, dihydrofolate reductase inhibitors, factor Xa inhibitors, farnesyl transferase inhibitors, matrix metalloproteinase inhibitors, and vitamin D analogues.

55. The medium as set forth in claim 50, wherein each bait compound has a particular biological activity.

56. The medium as set forth in claim 50, wherein said partitioning the compounds into partitions comprises:

dividing the compounds into a first partition of compounds which have the descriptor value greater than the median value and a second partition which have the descriptor value less than the median value.

57. The medium as set forth in claim 50, wherein the descriptor values are descriptor types independently selected from the group consisting of chemical properties, structural properties, surface area properties, and electrochemical properties.

58. The medium as set forth in claim 50, wherein the descriptor values are descriptor types independently selected from the group consisting of a sum of atomic polarizabilities of all atoms, a number of aromatic atoms, a number of H-bond donors, a number of heavy atoms, a number of hydrophobic atoms, a number of nitrogen atoms, a number of fluorine atoms, a number of sulfur atoms, a number of iodine atoms, a number of bonds between heavy atoms, a number of aromatic bonds, a number of double nonaromatic bonds, an atomic connectivity index (order 0), a carbon valence connectivity index (order 1), a carbon connectivity index (order 1), a greatest value in a distance matrix, a third kappa shape index, a relative negative partial charge, a total positive van der Waals surface area, a fractional negative polar van der Waals surface area, a fractional hydrophobic van der Waals surface area, a vertex adjacency information (magnitude), a vertex distance equality index, a vertex distance magnitude index, a sum of a van der Waals surface area of each of one or more atoms in each compound in the set of compounds, a van der Waals surface area calculated for a property of each compound selected from the group consisting of hydrogen-bond acceptor atoms, hydrogen-bond donor atoms, nondonor-acceptor atoms, and polar

atoms, a van der Waals volume calculated using a connection table, and a Zagreb index.

59. The medium as set forth in claim 50, wherein the descriptor values are different descriptor types that do not substantially correlate with each other.

60. The medium as set forth in claim 50 further comprising:

choosing different types of descriptors to base the descriptor values on using a genetic algorithm.

61. The medium as set forth in claim 60, wherein the different types of descriptors for the set of compounds each have value distributions from which the median values are calculated.

62. The medium as set forth in claim 60 further comprising:

establishing an optimal combination of the different types of descriptors to base the descriptor values on using the genetic algorithm.

63. The medium as set forth in claim 62, wherein a scoring function is used by the genetic algorithm during said establishing of the optimal combination of the different types of descriptors, the scoring function comprising:

$$S = Act(cp) \times Pa(pop),$$

wherein *Act(cp)* is a first total number of co-partitioned known active compounds in the set of compounds and *Pa(pop)* is a second total number of populated partitions.

64. The medium as set forth in claim 50, wherein said obtaining one or more descriptor values comprises:

calculating the descriptor values using a molecular modeling program.

65. A system for virtual compound screening comprising:

a bait compound system that combines information representing a plurality of unidentified compounds with information representing a plurality of bait compounds with known biological activities to form a set of compounds;

a descriptor system that obtains one or more descriptor values for each of the unidentified compounds and for each of the bait compounds in the set of compounds;

a median determination system that determines a median value for each of the descriptor values for the set of compounds;

a partitioning system that partitions the set of compounds into a plurality of partitions based on each median value;

a partition recombination system that recombines partitions which have at least two bait compounds to form a recombined set of compounds; and

a compound selection system that selects the recombined set of compounds for analysis of biological activity if an approximate target number of unidentified compounds remain in the recombined set of compounds.

66. The system as set forth in claim 65, wherein the compound selection system causes operation of the descriptor system, the median determination system, the partitioning system, and the partition recombination system with different descriptor values than used previously until the approximate target number of unidentified compounds remain in the recombined set of compounds.

67. The system as set forth in claim 66, wherein the compound selection system causes another set of bait compounds to be reintroduced into the recombined set of compounds substantially prior to operation of the descriptor system, the other set of bait compounds being identical to the bait compounds used by the bait compound system.

68. The system as set forth in claim 65, wherein the partitioning system divides the compounds into a first partition of compounds which have the descriptor value greater than the median value and a second partition which have the descriptor value less than the median value.

69. The system as set forth in claim 65, wherein the descriptor values are different descriptor types that do not substantially correlate with each other.

70. The system as set forth in claim 65, wherein the descriptor system chooses different types of descriptors to base the descriptor values on using a genetic algorithm.

71. The system as set forth in claim 70, wherein the different types of descriptors for the set of compounds each have value distributions from which the median values are calculated.

72. The system as set forth in claim 70, wherein the descriptor system establishes an optimal combination of the different types of descriptors to base the descriptor values on using the genetic algorithm.

73. The system as set forth in claim 72, wherein a scoring function is used by the genetic algorithm during establishment of the optimal combination of the different types of descriptors, the scoring function comprising:

$$S = Act(cp) \times Pa(pop),$$

wherein *Act(cp)* is a first total number of co-partitioned known active compounds in the set of compounds and *Pa(pop)* is a second total number of populated partitions.

74. The system as set forth in claim 65, wherein the descriptor system calculates the descriptor values using a molecular modeling program.